Clinical Oncology 34 (2022) e130–e134

Contents lists available at ScienceDirect

Clinical Oncology

journal homepage: www.clinicaloncologyonline.net

Editorial

Reading the Mind of a Machine: Hopes and Hypes of Artificial Intelligence for Clinical Oncology Imaging

A. Green [*†], M.C. Aznar [*†‡], R. Muirhead [§], E.M. Vasquez Osorio [*†]

[*] Radiotherapy Related Research Department, Division of Cancer Sciences, The University of Manchester, Manchester, UK
[†] Radiotherapy Related Research, The Christie NHS Foundation Trust, Manchester, UK
[‡] Clinical Trial Service Unit, Nuffield Department of Population Health, University of Oxford, UK
[§] Department of Oncology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

Artificial intelligence as a field was started when Alan Turing first described machines that could imitate humans and show intelligent behaviour [1]. Since then, artificial intelligence has grown into several sub-fields, one of them being machine learning, where a computer is trained to reproduce some output, mimicking a human. Even though big improvements have been seen in this field, some simple results show it is still challenging (Figure 1).

Throughout the 1970s, what we think of as modern machine learning began to be developed, with algorithms classifying data into several groups and the beginnings of simple neural networks [1]. For many years, machine learning was concerned with complicated feature [2] extraction techniques and, frankly, was not very good at tasks humans find trivial, for example saying whether there is a dog in an image. The results were a series of hypes, followed by disappointment and criticism, leading to the 'AI winter' [2], where research funding was drastically reduced and stopped substantial developments. This all changed in 2012, when a new paradigm emerged, known as deep learning [3].

In contrast to earlier attempts, deep learning neural networks include many 'layers' and need huge volumes of data to learn what features are most useful automatically. The most widely applied deep learning technique for imaging data is the convolutional neural network (CNN) [3], which requires very powerful computing hardware. Deep learning is now possible, and topical; because for the first time we have access to both large amounts of image data and powerful specialised computing hardware.

Due to the reliance on large datasets and automatic feature extraction, deep learning can be seen as a 'black box'. And as in any black box system, it is difficult to determine how the system works and follow the process taken to reach a conclusion. Interpretability matters, especially in healthcare, where a patient's care can be directly affected by these decisions. For example, if we were to use a machine learning model to diagnose patients with malignancy based on a medical image, we would need to understand how that conclusion is reached and how confident we are in that conclusion, in order to decide whether we deliver treatment without another form of verification.

The main applications in clinical oncology are summarised in Figure 2. The use of all of these applications requires some awareness of the process of deep learning and some method of quality assurance. As CNNs progressively become more abstract and complex, this need to interpret models and provide some 'quality assurance' to the output of CNNs becomes greater.

Looking Inside the Black Box: How Can We, as Users, Test Machine Learning Models?

An excellent comparison of interpretability can be made between two recent papers carrying out similar classifications of optical coherence tomography images of the retina. In this task, a machine learning model decides what disease is likely to be present in each image. In the paper by Kermany *et al.* [4], a classification was made directly from an optical coherence tomography image; we have no indication why that decision was made without extra post-processing. In contrast, in the paper by de Fauw *et al.* [5], a two-step process was used; first, the image was segmented, then the segmentation was classified. Even though the two-step process is slightly slower, the benefit is that the output of the first step can be visually checked, which may help diagnose wrong conclusions resulting from the second step. Neither approach is more correct, but the

Author for correspondence: A. Green, Radiotherapy Related Research Department, Division of Cancer Sciences, The University of Manchester, Manchester, UK.
E-mail address: andrew.green-2@manchester.ac.uk (A. Green).

https://doi.org/10.1016/j.clon.2021.11.008

Fig 1. Challenges of teaching a computer to mimic a human – how to define the difference between a chihuahua and a blueberry muffin.

two-step process is more interpretable. When using machine learning in a clinical setting, it is unlikely practitioners will have much influence over the design of the tool. However, we should be requesting information on interpretability when buying machine learning-based software and quality assuring the black box in ways that do not require direct access to the underlying model. We review two examples in the following.

Occlusion Mapping

Classification is the problem of identifying to which class a given image belongs. When a CNN classifies a given image, it considers information in small chunks of the image. This information is used to extract features, which are fed into a neural network to make a classification. Obviously, some regions of the image may imply different classes, for example wheels in an image could be on a car or a bicycle, but if handlebars are also present it is more likely to be a bicycle. To look at which regions of an image provided evidence for a given classification, we can use a technique called occlusion mapping. Occlusion mapping works by blocking small sections of the input image, running the model on this partially 'occluded' image and keeping track of the impact in the model's result. For example, to identify radiation pneumonitis, we can block out a part of the lungs, re-run the model and see whether the classification changes or remains the same. By doing this repeatedly, we can identify the area driving the classification. Occlusion mapping can be carried out completely independently of the underlying machine learning model – you just need a way to block out some of the image and run the model several times. In this way, independent quality assurance of a 'black box' classification model can be carried out without input from the vendor.

Dropout

Image segmentation is the problem of identifying the regions on the images containing certain objects, for example organs in a computed tomography scan. In medical images, the input is an image of a patient and the output is a series of masks covering anatomy of interest. As such, machine learning segmentation is already quite interpretable or at least easy to quality check – if the segmentation looks nothing like what a human/radiotherapy professional would draw, it is probably wrong. It is tempting to trust machine learning segmentation completely – many studies have been carried out to quantify the variation in segmentations between different human observers (known as inter-observer variability studies); surely a machine that is completely deterministic must be more trustworthy? Indeed, machine learning segmentations are more consistent, but they are still susceptible to uncertainty as they are trained on manually annotated data. This can be seen, for example, in the work by Kampffmeyer *et al.* [6], in which a measure of the uncertainty in a segmentation model is produced. To get this uncertainty, a technique called dropout is used in which different parts of the model (that identify different features in the image) are cut off at random and the model is re-run. This practically has the effect of giving a different segmentation each time, using a different subset of the model – similar to having multiple observers. In a robust model, this technique should not significantly affect the output. Unfortunately, in contrast to occlusion mapping, this type of testing requires access to the underlying model and will not be available for practitioners

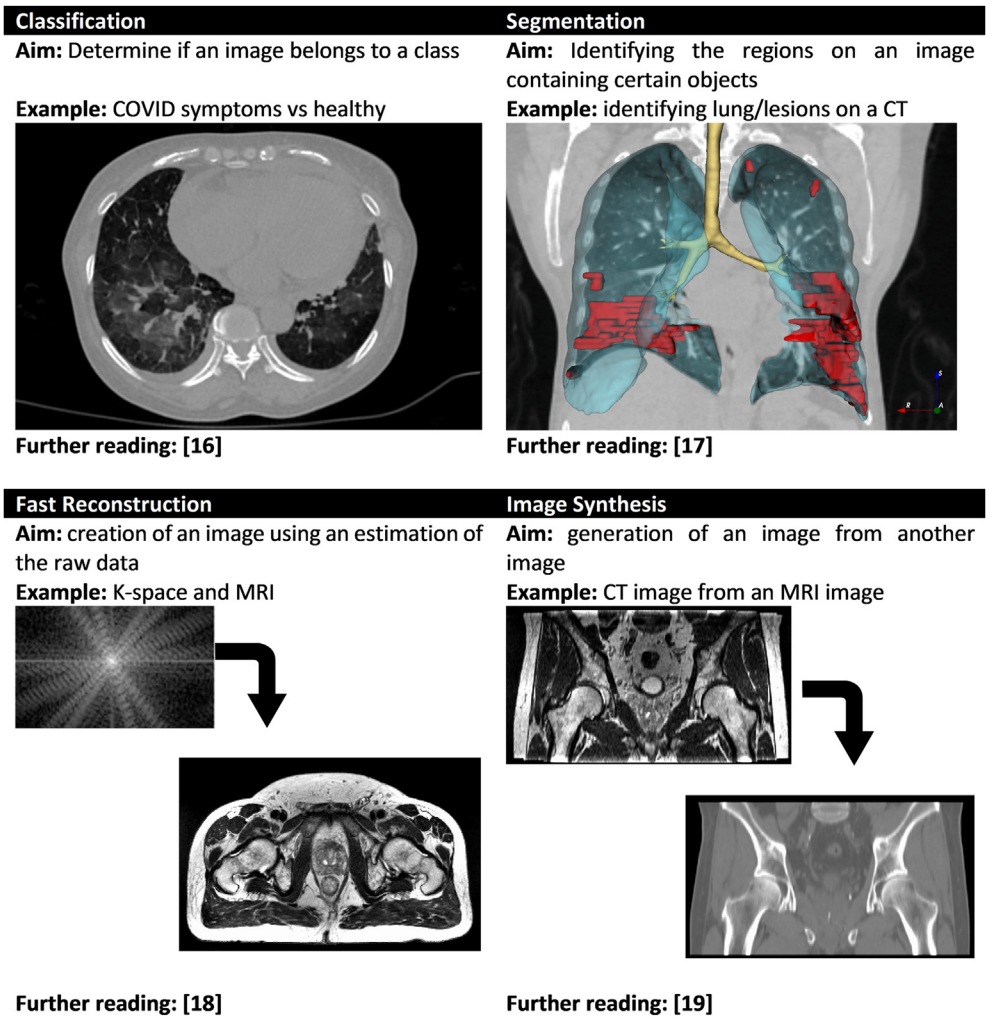

Fig 2. Examples of machine learning applied to images in clinical oncology. Example images sourced from open datasets [14,15]. Further reading [16–19].

using commercial products, as the model is owned by the vendor. This is, in our view, unacceptable, as a simple and effective means of uncertainty estimation exists: buyers should demand uncertainty quantification from vendors.

Further Potential Applications in Clinical Oncology: Image Generation

In addition to the use of classification in diagnostic radiology and automatic segmentation in radiotherapy planning, machine learning for image generation is also of particular interest to clinical oncology. Image generation is the problem of deriving an image based on a given input. The two areas of particular interest in clinical oncology are described below.

Fast Imaging

Machine learning may soon help to acquire and reconstruct images faster in order to get more information in less time. For example, for some magnetic resonance sequences, it can take over an hour for the final image to be ready, and it has been shown that by using machine learning techniques, this time can be reduced to less than 1 s without compromising image quality [7]. This relies on having access to the imaging raw data. Most images used in routine practice are not acquired in the form they are ultimately used. For example, computed tomography is acquired as a series of projections, whereas magnetic resonance is acquired as a path in k-space [4]: both types of 'raw data' are later reconstructed as three-dimensional images. Recently, work has begun to train machine learning models to work on the captured raw data and rapidly produce an 'estimated' volumetric image, bypassing conventional image reconstruction techniques [8,9]. These models promise faster imaging by reducing the amount of information required to be acquired to produce good images. Machine learning reconstruction model research is still in its infancy, and we are not aware of any attempts to introduce interpretability or quality assurance to these models.

Synthetic Images

Machine learning can be used to create one image from another. A recent area of growth in machine learning is the use of special CNNs to transfer the 'style' or appearance [10] from one image to another, for example taking a photograph and making it look like a van Gogh painting. This is done primarily through the use of generative adversarial networks (GANs) [11]. GANs work with a first network ('generator'), which tries to trick a second network ('discriminator'); this is an iterative process that continues until the generator is trained successfully. Successful training is achieved when the discriminator is not able to tell the difference between photograph-generated 'van Gogh' images and 'real' van Gogh-style images.

GANs are used in the context of medical imaging to generate synthetic images of one modality from another. One potentially exciting use for this is the generation of synthetic computed tomography from magnetic resonance images [12], allowing for magnetic resonance imaging-only radiotherapy. Given that these techniques can be trained to generate synthetic images of any modality, it may even be possible to derive one magnetic resonance sequence from another, although to our knowledge this has yet to be demonstrated.

Beware that synthetic medical images generated by machine learning should be treated with extreme caution. The GAN construct is not designed to create a physically meaningful mapping from one modality to another, only to be good enough to trick a discriminator into believing the image is 'of the same style'. As such, they will have no trouble inventing anatomy, or ignoring it entirely, leading to unreliable images. GANs are notoriously difficult to train and at present are uninterpretable.

Pathway to Clinic

Oncology represents a unique challenge for machine learning. As discussed, machine learning relies on large databases of data, allowing the computer to reach a conclusion using that background data. When looking at databases of 'normal' images, it is relatively easy to pick out the 'abnormal'. However, in oncology, we have innumerable versions of what cancers look like; as no one cancer is unique. Cancers are all different sizes, shapes, at different sites, with different densities and borders. These variations will never be fully represented in the training dataset.

In all types of machine learning there is the assumption that we have a suitable ground truth – an incontestable, true answer. In reality, this is almost never the case. The most egregious example in clinical oncology is that of segmentation. Inter-observer variation studies have repeatedly shown that no two clinicians will perfectly agree on where organ or tumour boundaries lie. When we build our dataset to train a model by collecting routine clinical examples (the most common approach), we are implicitly teaching the model that the boundaries of organs/tumours are poorly defined and should not be surprised when the resulting model sometimes produces contours we do not immediately agree with. Despite this, machine learning techniques will continue to develop, because machine learning on medical data presents potential solutions that will benefit the community as a whole.

Some technologies are closer to clinical use than others. Image segmentation is commercially available now, and can greatly reduce the time required to segment organs at risk during planning; caution is required in its use for tumour segmentation however, where there is still considerable room for interpretation by clinicians. To date, we are not aware of any commercial image segmentation tool that can provide an estimate of segmentation uncertainty – this is something manufacturers should include in the future.

Image generation techniques will begin to play a role in clinic soon, with synthetic computed tomography from magnetic resonance probably being the first application. When compared with the current standard for generating computed tomography-like images from magnetic resonance (i.e. density overrides based on segmentations), these may represent a significant improvement in dose calculation accuracy, provided the generated image can be properly checked and trusted. In theory, fast reconstruction is a better posed problem (we already know what the mapping from one space to the other should be) and therefore should be a more straightforward task. However, there remains the issue of how the algorithm will deal with abnormal tissue and whether it will produce a faithful reconstruction of the image.

Classification techniques are most useful early in the patient pathway, for example to diagnose disease. Classification is the easiest machine learning technique to quality assure, with several straightforward techniques available. There is enormous potential for classification to contribute to diagnostic radiology, particularly in view of the shortage of radiologists. For example, there are already ongoing trials in the use of machine learning to review screening mammograms to assess the safety of this technique, with a view to use in the clinic thereafter.

The pathway to clinic must include some consideration of the utility, time saving and cost-effectiveness of any introduced machine learning tool. While this discussion is beyond the scope of our editorial, we would encourage a review article on the subject, pulling together a multidisciplinary team with a strong focus on health economics. Such a multidisciplinary team will be key to elucidating the benefits of machine learning in the clinic, and where more work needs to be done. In the meantime, a technical overview of the implementation of machine learning in the clinic has been published, as a result of the 2019 ESTRO physics workshop on the subject [13]. Although this is primarily aimed at a physics audience, it discusses many of the potential problems faced during the commissioning of machine learning systems in clinic.

Given the unstoppable march of machine learning towards the clinic, now is the time to train clinicians to have a healthy scepticism of machine learning promises and to demand interpretability from vendors. These will be invaluable in the future as machine learning becomes more

involved in clinical work. Any tool intended for clinical use should be able to justify the influence it has on treatment and artificial intelligence tools should be no different. As users and developers of artificial intelligence we should demand and champion model interpretability as a basic requirement for clinical use.

Conflicts of Interest

The authors declare no conflict of interest.

Notes

1. A neural network is a series of algorithms that tries to recognise underlying relationships in a set of data through a process that mimics the way the human brain operates. It is one of the most common algorithms used in today's artificial intelligence.
2. Feature as in a distinctive attribute or aspect of something. For example, if a picture presents the face of a cat, it will probably show two pointy ears, two eyes, a triangular nose and some whiskers. These features can be extracted from images using image processing techniques.
3. Convolutional neural networks are neural networks that apply image processing techniques (convolutions) to images to find and extract features.
4. k-space is an array of numbers representing spatial frequencies in the magnetic resonance image. It is intuitively compared with an image of a 'galaxy', where each 'star' in k-space is just a data point derived directly from the magnetic resonance signal. The brightness of each star represents the relative contribution of that star's unique information to the final image.

Acknowledgements

This work was supported by Cancer Research UK via funding to the Cancer Research Manchester Centre (C147/A18083 and C147/A25254). All authors were also supported by the NIHR Manchester Biomedical Research Centre.

References

[1] Turing Alan. *Intelligent machinery (1948)*. B. Jack Copeland; 2004. p. 395.
[2] Hendler James. Avoiding another AI winter. *IEEE Intell Syst* 2008;23:2–4.
[3] Le Cun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–444. https://doi.org/10.1038/nature14539.
[4] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, *et al*. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122–1131. https://doi.org/10.1016/j.cell.2018.02.010. e9.
[5] De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–1350. https://doi.org/10.1038/s41591-018-0107-6.
[6] Kampffmeyer M, Salberg AB, Jenssen R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. *IEEE Comput Soc Conf Comput Vis Pattern Recog Work* 2016. https://doi.org/10.1109/CVPRW.2016.90.
[7] Mardani M, Gong E, Cheng JY, Vasanawala SS, Zaharchuk G, Xing L, *et al*. Deep generative adversarial neural networks for compressive sensing MRI. *IEEE Trans Med Imag* 2019;38:167–179. https://doi.org/10.1109/TMI.2018.2858752.
[8] Kang E, Min J, Ye JC. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Med Phys* 2017;44:e360–e375. https://doi.org/10.1002/mp.12344.
[9] Bao L, Ye F, Cai C, Wu J, Zeng K, van Zijl PCM, *et al*. Undersampled MR image reconstruction using an enhanced recursive residual network. *J Magn Reson* 2019;305:232–246. https://doi.org/10.1016/j.jmr.2019.07.020.
[10] Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recog* 2016;2016:2414–2423. https://doi.org/10.1109/CVPR.2016.265.
[11] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, *et al*. Generative adversarial nets. *Adv Neural Inf Process Syst* 2014;2014:2672–2680.
[12] Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Med Phys* 2018;45:3627–3636. https://doi.org/10.1002/mp.13047.
[13] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, *et al*. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. *Radiother Oncol* 2020;153:56–66. https://doi.org/10.1016/j.radonc.2020.09.008.
[14] An P, Xu S, Harmon SA, Turkbey EB, Sanford TH, Amalou A, *et al*. CT images in COVID-19. The Cancer Imaging Archive. Available at: https://doi.org/10.7937/TCIA.2020.GQRY-NC81.
[15] Nyholm T, Svensson S, Andersson S, Jonsson J, Sohlin M, Gustafsson C, *et al*. MR and CT data with multiobserver delineations of organs in the pelvic area - part of the Gold Atlas project. *Med Phys* 2018;45:1295–1300. https://doi.org/10.1002/mp.12748.
[16] Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z, *et al*. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nat Commun* 2020;11:4080. https://doi.org/10.1038/s41467-020-17971-2.
[17] Parkinson C, Matthams C, Foley K, Spezi E. Artificial intelligence in radiation oncology: A review of its current status and potential application for the radiotherapy workforce. *Radiography* 2021;27(Suppl):S63–S68.
[18] Lin DJ, Johnson PM, Knoll F, Lui YQ. Artificial intelligence for MR image reconstruction: an overview for clinicians. *J Magn Reson Imag* 2021;53:1015–1028.
[19] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, *et al*. GANs for medical image analysis. *Artif Intell Med* 2020;109:101938.